Songklanakarin J. Sci. Technol.
44 (5), 1365–1372, Sep. – Oct. 2022

http://www.sjst.psu.ac.th

Original Article

An innovative framework for extracting adverse drug reactions of single medication and combined medications from medical transcriptions and online reviews

Lakshmi K. S.*, and Liya Varghese

Department of Information Technology,
Rajagiri School of Engineering and Technology, Kochi, 682039 India

Received: 16 March 2022; Revised: 30 August 2022; Accepted: 15 September 2022

Abstract

Adverse drug reactions (ADRs) are unintentional and detrimental reactions arising due to normal drug usage. Identifying ADRs is vital in spheres of health and pharmacology. ADRs occur due to a single drug or a combination of multiple drugs. In the pharmaceutical industry, recognizing this type of medication interactions is viewed as a significant task. In this paper, we discuss the extraction of ADRs from combined medications (two drugs) by using medical transcripts and online reviews as the primary sources. Here, Natural Language Processing (NLP) techniques are combined with weighted association rule mining for extracting ADRs due to a single drug from medical transcripts. Single drug ADRs are also obtained from online health reviews using an ensemble classifier. These drugs along with their ADRs are used for constructing two-drug (combined medication) associated ADRs dataset. Further, by using the dataset of combined medications, the interaction of the medications and the reactions that are associated with that drug combination are predicted. In the first two phases of single drug associated ADR prediction, weighted association rule mining and ensemble classifier got an accuracy of 88%. The proposed model obtained an accuracy of 85.3%.

Keywords: ADRs, DDIs, text mining, weighted association rule mining, NLP, KeyBERT

1. Introduction

ADR is a major health issue everywhere on the planet. There are some studies exploring computational models to predict reactions to a single drug, but there is only very little research on reactions to combined drugs. Researchers have observed that, due to the interactions of drug combinations, approximately ten percent of plausible drug pairs can induce ADRs. Recent studies have revealed that there are no actual public databases with significant coverage of existing drug-drug interactions (DDIs). These may be incomplete or contain immense amounts of irrelevant interactions. Also, it is observed that a majority of DDIs are hidden within unstructured textual data. Medical Health Records, Clinical notes, and online reviews are the major sources of these unstructured data. With the advancements in technology, these resources can be wisely exploited for the discovery of ADRs due to single drug or even a combination of multiple drugs.

Ample research has been conducted over the past few decades for finding ADRs to single drugs. Statistical methods, machine learning methods and graphical methods were widely used for this purpose. Recently, novel techniques have been developed to aid the prediction of ADRs developed due to a combination of multiple drugs.

The objective of the proposed model is to construct a dataset having adverse events arising from a drug-pair that could be differentiated from any adverse event caused by the single constituents, and to predict adverse drug reaction from two-drug combination. For that, we applied Text Mining to medical transcripts and online healthcare forums such as “medications.com” and “steadyhealth.com” to produce a dataset for ADRs of specific pharmaceuticals by extracting the side-effects terms of that specific medication. From this

*Corresponding author
Email address: lakshmiks@rajagiritech.edu.in

dataset, based on some drug similarity measures (Cheng & Zhao, 2014; Ferdousi, Safdari, & Omidi, 2017; Kastrin, Ferk, & Leskošek, 2018; Zhang *et al.*, 2017) a dataset for ADRs of drug combination was created, which in turn was used to predict DDIs, resulting in ADRs for combined medications.

2. Related Work

Predicting potential ADRs in pre-clinical stages has become increasingly important lately, and varied machine learning-based models based on chemical features of compounds, knowledge of drug side effects, therapeutic indications, drug targets, enzymes, transporters, and pathways have been proposed. In most research, a predictive model is built on pre-defined structural traits, or fingerprints, initially obtained. Pre-defined chemical fingerprints, on the other hand, do not cover every possible chemical substructure, thus omitting vital matter. Some of the existing methods employed for Predicting Adverse Drug Reactions are discussed here.

Sampathkumar, Chen and Luo (2014) used a Hidden Markov Model based Text Mining system with an information retrieval module to collect information from healthcare forum messages, which was then processed using a text preprocessing module that included multiple Natural Language Processing tools for processing the text data. Then, by using Named Entity Recognition and sub-modules, the drug name, side-effect terminology and phrases to relate drug with its side-effect were identified. Relationship Extraction was treated as a sequence labeling problem.

For solving problems of data imbalance, a predictive model was developed by Jamal, Ali, Nagpal, Grover and Grover (2019) using random forest and sequential minimization optimization (SMO) to generate machine-learning based computational models for predicting cardiovascular related drugs' ADRs. The computational model was trained employing chemical, biological and phenotypic features and their two-three level combinations for 36 CV ADRs. For finding significant and non-redundant characteristics, the smote balancing method was employed, as well as minimum redundancy maximum relevance (mRMR) methodology.

Zhang, Sun, Diao, Zhao and Shu (2021) constructed a Knowledge Graph with four types of nodes representing the drugs, side effects, target indication and three relations that denoted the presence of the side effect, target and indication. These features were considered as characteristics of the drugs and used the Word2Vec model in Natural Language Processing to vectorize the graph. By using logistic regression, a binary classification model was implemented for evaluating if a side-effect relation exists between the drug and ADR.

A binary classifier was utilized to predict drug-drug-ADR associations from disparate pharmacologic databases by Zheng *et al.*, (2018). Diverse data were gathered and integrated from many databases, wherein every drug was represented in the form of a multi-dimensional vector to be utilized as classification inputs for calling ADR labels. For classification, any drug pair with a lower interaction probability was chosen as a negative sample. To measure its interaction probability, the drug-disease-gene graph was constructed, and the scoring method was devised to measure the interaction probability of all drug pairs via network analysis. Those causing the ADRs were taken as positive samples for the classification and were projected onto a lower-dimensional space utilizing principal component analysis, along with the binary classifier that was made for drug-drug-ADR association prediction.

For Inferring Drug–Drug Interactions, Gottlieb, Stein, Oron, Ruppin and Sharan (2012) developed a method that enables the prediction of drug interactions for new pharmaceuticals, wherein interaction information available is nil, by using the new prediction method called INDI (INferring Drug Interactions). This was based on Chemical, Ligand, Side-effect, Annotation and Sequence. Drug-drug similarity metrics were used for identifying drug-drug interactions with 93% sensitivity and specificity.

A model was built by Li, Tong, Zhu and Zhang (2020) using machine learning for drug combination prediction by utilizing multi-feature data on drugs. For determining similarity of drug pairs, Tanimoto coefficient was utilized, and the neighbor-recommender method was paired alongside ensemble learning algorithm to increase prediction accuracy. A feature assessment study was also performed to pick the most useful drug attributes, and the ensemble models attained an AUC of 0.964.

Mahadevan, Vishnuvajjala, Dosi, and Rao (2019) adopted a predictive model based on similarity-based ensemble prediction for identifying potential DDIs of approved drugs and unapproved drugs. Eight features, obtained from five different databases, and known interactions helped characterize any drug-drug interaction. They calculated the similarity score between the drugs using Jaccard's coefficient and used these similarity values in the random walk method and neighbor recommender method for DDI prediction. These models with ensemble rules (classifier and weighted average) were then congregated to develop the ensemble model that could achieve superior performance. They further improved the prediction model by using genetic algorithm techniques.

A new feature-based device was proposed to identify ADRs from social media by Zhang, Cui and Gao (2020). Twitter and DailyStrength data were collected, and with the help of domain experts, DailyStrength data were annotated. For each user post, guided by extended syntactic dependencies and ADR lexicon, predicate-ADR pairs were extracted and by extended FactNet knowledge base and other domain knowledge, the POS and semantic features for the pairs were extracted for creating an initial deep feature set. For holistically representing deep linguistic features, these initial features of different pairs were pooled, and finally, the deep linguistic features were combined with many shallow linguistic features for training the predictive model.

3. Materials and Methods

This paper presents a method for predicting ADRs of single medication and combined medications by using Text Mining and the drug-drug interactions semantic similarity features using statistical learning. Figure 1 shows the overall architecture of the proposed system.

For ease of understanding, the proposed model has been divided into 3 modules. The first module deals with the creation of dataset for ADRs of individual drugs by extracting ADRs from medical transcription by using weighted

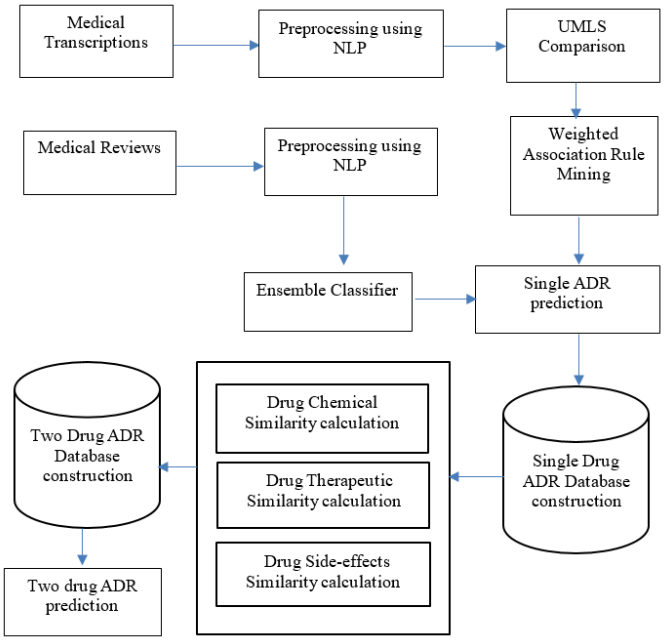

Figure 1. Proposed architecture

association rule mining. The second module deals with the extraction of ADRs from medical reviews using tf-idf weighting scheme and KeyBERT for keywords and phrases extraction and an ensemble classifier for ADR prediction. The third module is for detecting the drug-drug interactions of pairs of drugs and for displaying the ADRs of two drug combinations.

3.1 Extracting ADRs from medical transcriptions

Medical transcription (MT) manually processes voice reports of physicians and healthcare professionals into textual content. Healthcare providers record their notes which must be converted into text, usually in digital format. From these medical transcripts, medical terms as well as phrases were extracted using Natural Language Processing (NLP) techniques. Figure 2 Shows the extraction of ADRS from medical transcriptions.

Six main stages in extraction of medical terms are:

i. Sentence extraction (parsing) and stop word removal,
ii. POS tagging,
iii. UMLS comparison,
iv. Stemming,
v. Synonym finding
vi. Negative scope identification

The first phase to extract medical terms is by parsing the sentences from these transcripts. Syntax in a sentence was created by Stanford parser. For every sentence, the dependency tree was made to retrieve syntactical information. Most terms are multi-worded phrases. POS (part of speech) tagger was employed to extract these medical terms. With the help of tagger, consecutively noun tagged words, adjective tagged followed by noun tagged words, verb, gerund or present participle tagged words, besides noun, verb and adjective phrases were collected. Stop words were put in another file. Next, the stop words were eliminated through comparison of contents in stop word file. These terms from dependency tree were explored in UMLS, a medical ontology for segregation of terms. Here, the customized UMLS was installed utilizing metamorphosis, which incorporates an immense set of source vocabularies defining ample medical terms. Almost the entire vocabulary in UMLS has been chosen. These contain the terms as database tables. For enabling easy manipulations, UMLS database tables were loaded into relation tables in MySql. The UMLS contains MRCONSO data file with a term and its Concept Unique Identifier (CUI). The CUI helps find the semantic type of relevant terms, retrieved from MRSTY table. Mostly, there could be more than a single semantic type for one CUI. For avoiding this problem, semantic types given in more vocabularies for CUI were considered. Synonyms of corresponding terms can aid UMLS comparison. If a term is not present in UMLS, its synonym could be got through Wordnet, and checking can be done in UMLS. After extraction of terms, these were included in XML file to create frequent patterns. Stemming could be carried out before UMLS comparison to improve exactness. Terms stated in negative connotation must be removed from mining phase for which NegEx algorithms by Chapman, Bridewell, Hanbury, Cooper and Buchanan (2001) could be employed. After the extraction of medical terms, association rule mining was done to extract drug-ADR relationships. The pseudocode given is used for weighted association rule mining algorithm by Yun, Russel, and Wai (2010). In this algorithm, weight is taken as the valency value of the term, where valency is given by the following equation:

$$v_k = \beta \cdot p_k = n\beta(1-\beta) \cdot p_k \quad (1)$$

where

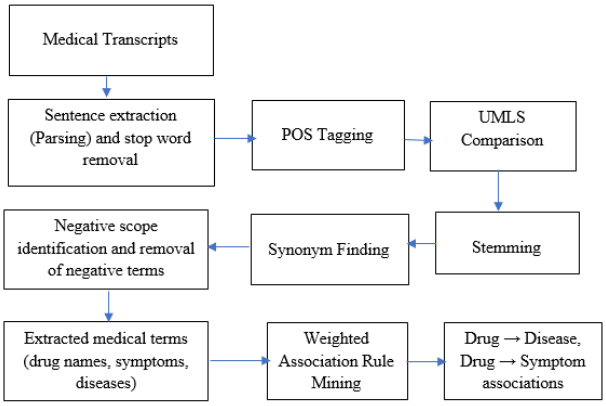

Figure 2. Extracting ADRs from medical transcriptions

$$p_k = 1 - \frac{\log_2(|I_k|) + \log_2(|I_k|)^2}{\log_2(|U|)^3} \quad (2)$$

$$\beta = \frac{1}{n}\sum_{i}^{n}\frac{count(ik)}{count(k)} \quad (3)$$

In equation (2), |U| denotes number of unique items in dataset and $|I_k|$ denotes those co-occurring with item k. In equation (3), n denotes total number of transactions.

Using the weighted association rule mining, rules of the form medicine → symptoms and medicine → disease were extracted. The RHS of the rules correspond to the ADR of drug mentioned in the LHS. These ADRs and their respective drugs are added to the ADR file.

Algorithm: Weighted Association Rule Mining (WARM)

Input: Transaction database *D*, weighted minimum support *wminsup*, universe of items *I*

Output: Weighted Frequent itemsets

$L_k \leftarrow \{\{i\} | i \in I, \text{weight}(c) * \text{support}(c) > \text{wminsup}\}$

$k \leftarrow 1$

while ($|L_k| > 0$) do

$k \leftarrow k + 1$

$C_k \leftarrow \{x \cup y \,|\, x, y \in L_{k-1}, |x \cap y| = k-2\}$

$L_k \leftarrow \{c | c\ C_k, \text{weight (c) * support (c)} > wminsup\}$

$L_k \leftarrow \bigcup_k L_k$

3.2 Extracting ADRs from medical reviews using web scraping

Web scraping method involves the fetching and extraction of information from a website. For the creation of a single drug ADR dataset, we collected information or reviews about the drugs from an online healthcare forum, such as "medications.com" by using this web scraping method.

Medications.com is an online forum for having discussions regarding the drugs, conditions and other information pertaining to medications. It contains posts or reviews relating to thousands of drugs from different groups of populations, which offer a perfect source to extract the drug and its side effects. Here, each drug has its own discussion board; wherein there are multiple threads formed to talk on drug specifics. Figure 2 shows a screenshot of sample messages posted on the forum discussion of the drug *Synthroid* on the website "medications.com". In total 11,235 posts were collected from medications.com across 1,750 threads. Figure 3 shows sample messages posted on "medications.com".

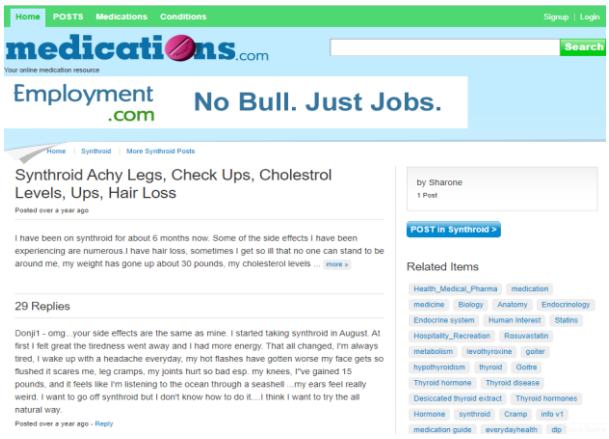

Figure 3. Sample messages posted on "medications.com"

Reviews were also collected from www.steadyhealth.com. 32,540 posts across 12,340 threads were collected from this site. As done in the HMM model, proposed by Sampathkumar *et al.*, (2014), dataset from medications.com was utilized to train and the dataset collected from steadyhealth.com was used for testing.

After obtaining the data through web scraping, the next step was text preprocessing, which is to clean and transform textual data into a usable form. It was performed on this web-scraped content to extract relevant information about the drugs.

For keyword extraction, we implemented two techniques. The first one was the Keywords Extraction with TF-IDF and the second was the Keywords Extraction with BERT. TF-IDF, or Term Frequency – Inverse Document Frequency is a vital approach to represent how relevant a word or phrase is to a particular document for information retrieval. One of the models for natural language processing is through transformers and is called Bidirectional Encoder Representations from Transformers (BERT). KeyBERT extracts keywords from a text using the BERT approach. KeyBERT is a simple keyword extraction technique that uses BERT embeddings as well as simple cosine similarity to locate document sub-sections. On web-scraped content, keyword extraction was done, and the outcomes of both strategies were compared. It has been found that KeyBERT has a better performance in keyword extraction from given data when compared to the first alternative above. So, the proposed system proceeded with the KeyBERT technique. Figure 4 gives the descriptions of second module.

Prediction of ADR was done in three phases. The first phase dealt with the extraction of Named Entities. The second phase extracted relationship features. The third phase was the classification phase using an ensemble classifier. After extraction of the named entities and keywords, they were compared with UMLS for identifying drugs, symptoms and diseases. In the relationship extraction phase, relationships between the named entities were identified. 45 phrases and keywords were selected to extract relationships. Following are the lists of keywords and phrases pointing to

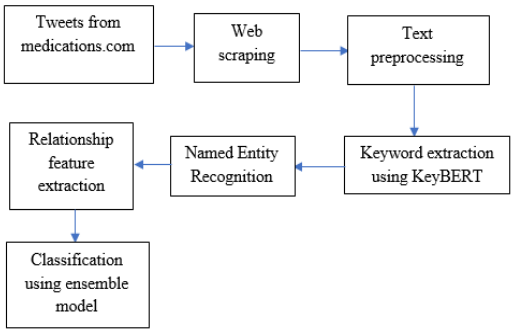

Figure 4. Extracting ADRs from medical reviews using web scraping

the relation between a drug and its ADR: {'after having', 'after stopping', 'because of this', 'caused by', 'cause of', 'developed', 'due to', 'effects from', 'effects of', 'ever since', 'experienced', 'experiencing', 'feeling', 'feel like', 'felt like', 'found out', 'found that', 'had a problem', 'have been getting', 'have been having', 'have noticed', 'have started', 'I am having', 'I am starting', 'I now have', 'Made me feel', 'Makes me feel', 'Now I have', 'Problem with', 'Problems with', 'Reaction to', 'result to', 'side affects', 'side effect', 'side effects', 'since I got', 'since I stopped', 'since then', 'started getting', 'started having', 'started noticing', 'started taking', 'started to', 'starting to feel', 'was causing'}(Sampathkumar *et al.*, 2014).

An ensemble classifier was used for learning the relationships between drugs and the side effects. The machine learning algorithms utilized in ensemble classifier were SVM, HMM, and Neural Network classifiers. The dataset comprised a training set and a test set. After training the classifiers with the training set, they were tested using the test set. Maximum accuracy was found for the ensemble classifier, superior to the individual machine learning algorithms on our dataset, and hence it was chosen as the backend algorithm for this prediction.

After extracting drugs and ADRs from medical transcripts and online reviews in the first two phases, a dataset was constructed for storing this single drug ADR relationships. Now, Figure 5 gives the description about Drug Interaction Module.

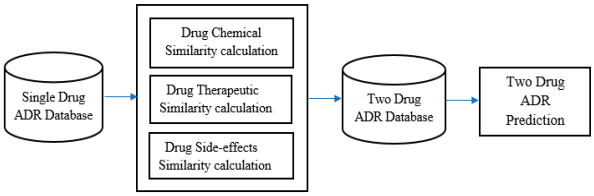

Figure 5. Flow of the third module

3.3 Drug similarity calculations

For finding potential drug-drug interactions, we rely on the drug-drug similarity measures. There are various alternative similarity measures like Jaccard's coefficient, Euclidean distance etc. To measure the similarity between two drugs the metrics used were:

Chemical-based: For the chemical-based drug similarity calculation, drug molecules were obtained from PubChemPy, which provides a way to interact with PubChem in Python. The similarity of two molecules was calculated using molecular fingerprints, which recorded structural information about the molecule as a series of 0s and 1s that signify the existence or absence of substructures. Two molecules that contain more of the same patterns will have more bits in common, which indicates that they are more similar. In most cases, the Tanimoto coefficient was used to determine whether two fingerprints were similar (Holliday, Salim, Whittle, & Willett, 2003; Vilar et al., 2012).

Annotation-based: For annotation-based similarity calculations, Anatomical Therapeutic Chemical (ATC) Classification System was used. An ATC code was assigned to each drug according to its actions, the effect it has on an organ or system, and its chemical characteristic (Kastrin & Leskošek, 2018). These codes were gained from DrugBank. To define similarities among ATC terms, normalized Hamming distance was used. Since the hamming distance is giving the dissimilarity between the strings in a range of 0 to 1, we obtained the similarity value by subtracting the Normalized Hamming distance by one.

Side-effect based: The drug side effect was gained from the dataset that we have created using text mining methods. The similarity between the two drugs was calculated with Jaccard score as mentioned in Sridhar, Fakhraei and Getoor (2016).

3.4 Two drug dataset creation

When the similarities for all the drugs in the single drug ADR dataset were obtained, then the dataset was created of ADRs for combined medications. This dataset will be storing the common ADRs and the similarity scores of all possible drug combinations, which can be made from the single drug ADR dataset. So, when a new drug or drug combination is given to the model, such that was not present in both the datasets created, then model will perform its entire actions to get their data for appending it to the single drug ADR dataset, which was then used for calculating the similarities and for detecting drug-drug interactions when drugs are administered together.

3.5 Detecting drug-drug interactions

For detecting a drug-drug interaction, the model initially set a threshold value. This threshold was taken as the average of the total similarities of the drug chemical structure similarity and the drug therapeutic similarity. For each possible drug combination, the model will check whether the total similarity value of the drug combination was greater than or equal to the threshold set. If its total score was greater than the average score, then drugs were interacting, and it was appended to a drug interaction file which will be storing all the possible drug combinations that have interactions.

3.6 Extracting adverse drug reactions for two-drug combination

When a drug-pair is given to the model, it will check this drug interaction file. If the drug combination is

found in the drug interaction file, then it means that there is a chance that the drugs interact and the model then displays the ADRs of the drug combination. If the drug combination is not found, it will then check for the drugs in the single drug ADR dataset, if both the drugs are present in the single ADR dataset then the model will display the ADRs of individual drugs. If both the drugs or one of the drugs from a combination are not there in our single drug ADR dataset, then the model will extract the review of that particular drug whose ADRs are not present in the database from the forum, perform keyword extraction, filtering, and appends the drug names and their corresponding ADRs to the single drug ADR dataset, calculates similarity measure, checks for interaction and stores them in the drug interaction file, if there is interaction, and then displays the ADRs.

4. Results and Discussion

Here, the result from each proposed system and the combined results are discussed. Table 1 shows the performance of the weighted association rule mining of ADRs from medical transcripts. Figure 6 outlines correlation of confidence values against number of rules generated, which demonstrates the number of times they are viewed as authentic. Thus, higher confidence suggests higher strength of a given association rule. In the graph, y axis is in the scale of 1:100.

For creating the dataset for ADRs of individual drugs from medications.com we have implemented 2 methods of Text Mining and proceeded with the KeyBERT method since it has better performance among alternatives on the data that is being collected.

After collecting named entities, keywords and phrases, three classifiers (SVM, HMM and Neural Network) were used for prediction. Based on a voting scheme, the prediction with highest number of votes was chosen and added to the final set of ADRs along with the drug name as shown in Table 2. Table 3 shows the results of 10-fold cross-validation run for the ensemble classifier, and Table 4 shows the performance metrics of the ensemble classifier.

Table 1. Performance of weighted association rule mining algorithm

Performance metric	Score (%)
Accuracy	88
Precision	80
Recall	89
F1 score	84

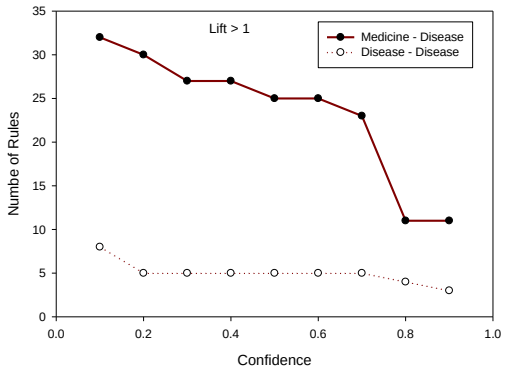

Figure 6. Comparison of confidence values against number of rules generated

For the creation of dataset for ADRs of combined medications, we calculate the Chemical, Therapeutic and Side-effects based similarities between two drugs in the single drug ADR dataset and a dataset is created with these similarity values as attributes. From this new dataset based on the threshold values, drug combinations having ADRs are detected and displayed. Here, Table 5 shows the drug interactions obtained for 5 drug combinations and Figure 7 shows the implementation result of the proposed system for a given drug combination.

Observed drug interactions are validated against expected interactions obtained from the DrugBank site. The implementation of the proposed models was assessed utilizing

Table 2. Top drug – ADR associations obtained for single drugs from the ADR dataset

Drug	ADRs
Acetaminophen	Orthovisc Body ache, Lateral Side, Excruciating Pain, Big Toe
Adderall	Pain in bladder, Pounding, irregular heartbeat or pulse, Anxiety, Dry mouth, Cloudy or bloody urine, Frequent urges to urinate, Pain in lower back or side, Stomach ache, Weight loss, Lack of strength
Amlodipine	lips and tongue swelled, canker sores, horribly depressed, Swollen Feet and Ankles, Tingling and Pain, swollen legs and feet, dry persistent cough, severe heartburn, stomach burns
Amoxicillin	Depressed and verging on suicidal, stomach cramping, twisting pain extreme lethargy and sleepiness, body aches, headaches and nauseous, diarrhea, body feels itchy, swollen neck, Short of Breath, Panic attack
Ativan	Nerve pain, tingling sensation, bruises, bone and muscle pain with terrible muscle twitching, heart palpitations, withdrawal symptoms, Extreme Anxiety, Deep Breathing
atorvastatin/Lipitor	Achy joints and lower back pain, soreness in wrists, thigh pain, Loss of vision, hearing, taste, smell, Fibromyalgia with severe chronic pain, nerve pain, sore muscles, and problems of concentration
benzonatate	Jittery and nervous, black stool, headaches, choking feeling, numbness, panic attack and heart palpitations, severe heartburn
Norvasc	swollen legs and feet, Tingling & Pain, severe cough, swollen ankles, sore gums, bleeding gums, memory loss
Lisinopril	Dry cough, severe body pains, weakness in arms and legs, vision impairment, hair loss, cough, sore throat, severe dry mouth, blistering on the back
Zyrtec	joints swollen, Lower back pain, tiredness, pains in arms, diarrhea
Singulair	nightmares and panic attacks, lost bladder control, overweight, terrible mood swings and depression

Table 3. Output of ensemble classifier

Iteration	Train set	Test set	True positive	False positive	True negative	False negative	Precision	Recall	Accuracy	F-score
1	1900	230	51	12	141	26	0.81	0.66	0.83	0.73
2	1900	230	42	9	152	27	0.82	0.61	0.84	0.7
3	1900	230	61	10	135	24	0.86	0.72	0.85	0.78
4	1900	230	57	8	143	22	0.88	0.72	0.87	0.79
5	1900	230	44	9	161	16	0.83	0.73	0.89	0.78
6	1900	230	56	9	144	21	0.86	0.73	0.87	0.79
7	1900	230	59	10	146	15	0.86	0.8	0.89	0.83
8	1900	230	48	12	147	23	0.8	0.68	0.85	0.74
9	1900	230	55	11	151	13	0.83	0.81	0.9	0.82
10	1900	230	57	10	149	14	0.85	0.8	0.9	0.82

Table 5. Two-drug ADR dataset for 5 drug combinations

Drug-1	Drug-2	ADRs
Topamax	Kenalog	migraine, pneumonia, heartburn, cancer, herniation, epilepsy, diarrhea, depressive, diabetic
Topamax	Wellbutrin	migraine, diarrhea, heartburn, cancer, depressive, diabetic
Topamax	Synthroid	migraine, heartburn, nightmares, diabetes, fainting, diarrhea
Avelox	Levaquin	heartburn, pneumonia, nightmares, diabetes, hypertension, colitis, diarrhea
Avelox	Lisinopril	rheumatoid, diarrhea, pneumonia, heartburn, nightmares, diabetes, hypertension, debilitating, arthritis, cancer

Table 4. Performance of ensemble classifier

Performance metric	Score (%)
Accuracy	88
Precision	83
Recall	72
F1 score	78

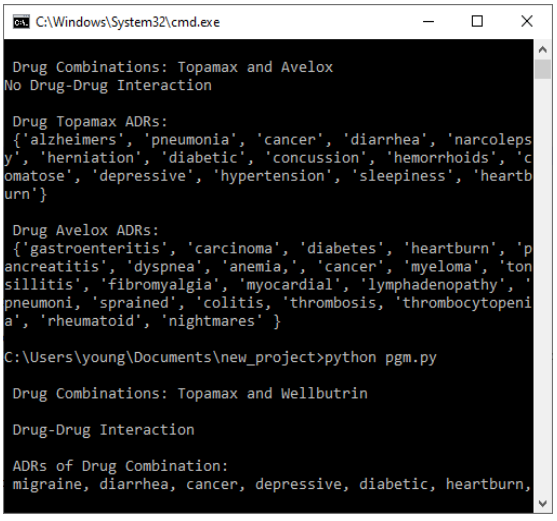

Figure 7. Implementation result

precision, accuracy, recall and F1-measure. The model got an accuracy of 85%, precision of 86%, recall of 72%, and F1-score of 0.78 from detecting the drug interactions for 145 drug combinations. Table 6 shows the evaluation of the performance metrics for the proposed system.

Table 6. Performance analysis of the proposed model

Performance metric	Score (%)
Accuracy	85
Precision	86
Recall	72
F1 score	78

5. Conclusions and Future Work

The detection and prediction of ADRs are paramount in drug safety surveillance. There are some existing methods for detecting associations from related records, which rely on expensive wet-lab experiments. Here, we were using data from an online healthcare forum where the information is of varied quality. It contains a huge amount of content as the patient numbers keep increasing. The first module created the dataset for ADRs of single drugs by extracting side-effects from the online healthcare forum. This was then followed by the creation of a dataset for combined medications from which we detect the DDI of the paired medications. The model was implemented in such a way that it displays the ADRs of combined medications, if the drugs when administered together are interacting. Otherwise it will display the ADRs of the individual drugs separately. This model can be further improved by adding more features for finding similarities, so that we get more accurate interactions among the various drug combinations.

References

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical*

Informatics, *34*(5), 301–310. doi:10.1006/jbin.2001.1029

Cheng, F., & Zhao, Z. (2014). Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association: JAMIA*, *21*(e2), e278–e286. doi:10.1136/amiajnl-2013-002512

Ferdousi, R., Safdari, R., & Omidi, Y. (2017). Computational prediction of drug-drug interactions based on drugs functional similarities. *Journal of Biomedical Informatics*, *70*, 54–64. doi:10.1016/j.jbi.2017.04.021

Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E., & Sharan, R. (2012). INDI: A computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology*, *8*, 592. doi:10.1038/msb.2012.26

Holliday, J. D., Salim, N., Whittle, M., & Willett, P. (2003). Analysis and display of the size dependence of chemical similarity coefficients. *Journal of Chemical Information and Computer Sciences*, *43*(3), 819–828. doi:10.1021/ci034001x

Jamal, S., Ali, W., Nagpal, P., Grover, S., & Grover, A. (2019). Computational models for the prediction of adverse cardiovascular drug reactions. *Journal of Translational Medicine*, *17*(1), 171. doi:10.1186/s12967-019-1918-z

Kastrin, A., Ferk, P., & Leskošek, B. (2018). Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. *PloS one*, *13*(5), e0196865. doi:10.1371/journal.pone.0196865

Li, J., Tong, X. Y., Zhu, L. D., & Zhang, H. Y. (2020). A machine learning method for drug combination prediction. *Frontiers in Genetics*, *11*, 1000. doi:10.3389/fgene.2020.01000

Mahadevan, A. A., Vishnuvajjala, A., Dosi, N. & Rao, S. (2019). A predictive model for drug-drug interaction using a similarity measure. *Proceeding of the 16th IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 1–8. doi:10.1109/CIBCB.2019.8791458.

Sampathkumar, H., Chen, X. & Luo, B. (2014) Mining adverse drug reactions from online healthcare forums using Hidden Markov Model. *BMC Medical Informatics and Decision Making*, *14*(1), 91. doi:10.1186/1472-6947-14-91

Sridhar, D., Fakhraei, S., & Getoor, L. (2016). A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics (Oxford, England)*, *32*(20), 3175–3182. doi:10.1093/bioinformatics/btw342

Vilar, S., Harpaz, R., Uriarte, E., Santana, L., Rabadan, R., & Friedman, C. (2012). Drug-drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association: JAMIA*, *19*(6), 1066–1074. doi:10.1136/amiajnl-2012-000935

Yun Sing Koh, Russel Pears, & Wai Yeap, (2010). Valency based weighted association rule mining. In Zaki M. J., Yu J. X., Ravindran B., Pudi V. (Eds.), Advances in knowledge discovery and data mining. PAKDD. *Lecture Notes in Computer Science, Volume 6118*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-13657-3_31

Zhang, F., Sun, B., Diao, X., Zhao, W., & Shu, T. (2021). Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Medical Informatics and Decision Making*, *21*(1), 38. doi:10.1186/s12911-021-01402-3

Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., & Li, X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics*, *18*(1), 18. doi:10.1186/s12859-016-1415-9

Zhang, Y., Cui, S., & Gao, H. (2020). Adverse drug reaction detection on social media with deep linguistic features. *Journal of Biomedical Informatics*, *106*, 103437. doi:10.1016/j.jbi.2020.103437

Zheng, Y., Peng, H., Zhang, X., Zhao, Z., Yin, J., & Li, J. (2018). Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases. *BMC bioinformatics*, *19* (Supplement 19), 517. doi:10.1186/s12859-018-2520-8